Biodiversity Data Journal 7: e28737
doi: 10.3897/BDJ.7.e28737

open access

Software Description

Biodiversity Observations Miner: A web application to unlock primary biodiversity data from published literature

Gabriel Muñoz[‡,§], W. Daniel Kissling[|], E. Emiel van Loon[|]

‡ NASUA, Biodiversity research and conservation section, Quito, Ecuador
§ Faculty of Arts and Science, Department of Biology, Concordia University, Montreal, Canada
| Faculty of Science, Institute for Biodiversity and Ecosystem Dynamics, University of Amsterdam, Amsterdam, Netherlands

Corresponding author: Gabriel Muñoz (nasua.research@gmail.com)

Academic editor: Donat Agosti

Received: 30 Jul 2018 | Accepted: 19 Dec 2018 | Published: 16 Jan 2019

Citation: Muñoz G, Kissling W, van Loon E (2019) Biodiversity Observations Miner: A web application to unlock primary biodiversity data from published literature. Biodiversity Data Journal 7: e28737. https://doi.org/10.3897/BDJ.7.e28737

Abstract

Background

A considerable portion of primary biodiversity data is digitally locked inside published literature which is often stored as pdf files. Large-scale approaches to biodiversity science could benefit from retrieving this information and making it digitally accessible and machine-readable. Nonetheless, the amount and diversity of digitally published literature pose many challenges for knowledge discovery and retrieval. Text mining has been extensively used for data discovery tasks in large quantities of documents. However, text mining approaches for knowledge discovery and retrieval have been limited in biodiversity science compared to other disciplines.

New information

Here, we present a novel, open source text mining tool, the **Biodiversity Observations Miner (BOM).** This web application, written in R, allows the semi-automated discovery of punctual biodiversity observations (e.g. biotic interactions, functional or behavioural traits and natural history descriptions) associated with the scientific names present inside a corpus of scientific literature. Furthermore, BOM enable users the rapid screening of large quantities of literature based on word co-occurrences that match custom biodiversity dictionaries. This tool aims to increase the digital mobilisation of primary biodiversity data and is freely accessible via GitHub or through a web server.

Keywords

biodiversity data, biodiversity knowledge, biotic interactions, data mobilisation, scientific names, text mining, R.

Introduction

Mobilisation, digitalization and interoperability of data on biodiversity are vital for sharing our global knowledge of nature (Wilkinson et al. 2016, Kissling et al. 2015, Edwards 2000). The need for digitally available biodiversity data has resulted in the development of global cyber-infrastructures such as the Global Biodiversity Information Facility (GBIF: www.gbif.org) (Edwards 2001), the Plant Trait Database (TRY: www.try-db.org) (Kattge et al. 2011), the Data Observation Network for Earth (DataOne: www.dataone.org) (Michener et al. 2011) and Global Biotic Interactions (GloBi: www.globalbioticinteractions.org) (Poelen et al. 2014). Those efforts have made digital biodiversity data increasingly available in recent years. However, a considerable amount of biodiversity data is still locked inside the current corpus of published literature (Nguyen et al. 2017). This pool of biodiversity data is often stored and shared as PDF files which limits its interoperability. With the increasing availability of literature on the internet, unlocking this biodiversity data and making it digitally interoperable becomes a challenge. Hence, there is a need for developing automatic and semi-automatic computational tools to discover and mobilise biodiversity data contained within this large corpus of literature (Senderov et al. 2017).

Text mining is a computational technique used for the automatic and semi-automatic discovery of useful information from large quantities of text (Hearst 2012). In bio-medicine research, text mining is applied for time-demanding tasks such as document classification and for the discovery of novel potential protein functions and protein-protein interactions (Petrič and Cestnik 2014, Saffer and Burnett 2014, Tari and Patel 2014). Biodiversity data stored within literature can be found in scientific articles (Thessen et al. 2012) or books and monographs (Kissling et al. 2014a). Recently, Algorithms and Application Programmatic Interfaces (APIs) have been developed for the recognition of taxonomic entities and

semantic tagging of ecological literature (Nunez-Mir et al. 2016, Pyle 2016, Sautter et al. 2006, Thessen et al. 2012). Furthermore, as ecology moves towards a data-driven science (Michener and Jones 2012), interest in the use of text mining frameworks for data discovery is growing (Miller et al. 2012,Thessen et al. 2012, Thessen and Parr 2014, Nunez-Mir et al. 2016,Senderov et al. 2017, Nguyen et al. 2017,Senderov et al. 2018).

Here, we present the **Biodiversity Observations Miner (BOM)**, a text mining tool that has been designed to augment the ability of ecologists and biodiversity scientists to implement text mining frameworks into their data compilation workflows. A first approach of implementing BOM into biodiversity research is using it as a tool to speed up and standardise the selection of candidate articles for large-scale meta-analyses. In addition, BOM can also be used for rapid discovery of specific biodiversity data across multiple articles at once. As such, this web tool can be used to discover observations from literature and to populate global biodiversity databases, for example on species traits (e.g. TRY) or species interactions (e.g. GloBI). As such, the BOM allows increasing the digital accessibility and availability of biodiversity data. The main feature of BOM is to identify snippets of text that potentially contain biodiversity information (i.e. data of biodiversity observations) within a given corpus of literature. BOM finds these snippets either by finding text statements linked to taxonomic entities (e.g. species names, genus, family) or by using specific keywords to filter a rank of annotated word co-occurrences inside the corpus of literature. These keywords are a curated list of terms describing a particular biodiversity observation and are provided in BOM as biodiversity dictionaries. Biodiversity Observations Miner is open source and freely accessible via GitHub (BiodiversityObservationsMiner) or via a web server (goo.gl/wt6V9R).

Project description

Design description: User interface:

The web application follows a dashboard design containing a header, a sidebar menu and the main page (Fig. 1). The dashboard header is placed at the top of the screen where users can find the application name (i.e. Biodiversity Observations Miner), a button to collapse the sidebar menu and a notification menu. The sidebar menu is located at the left side and allows easy navigation across all the specific functionalities of BOM. Clicking on each of the tabs in the sidebar menu will render a different content in the main page. The main output of the BOM consists of a list of text snippets, each a sentence long, indexed and annotated across all literature uploaded to the application. Thus, a user can perform a rapid literature search by filtering the output snippets based on taxonomic content (using scientific names present in the text) or biodiversity dictionaries (using curated lists of biodiversity terms). In addition, the application provides an overview of the semantic context of text snippets by calculating patterns of word co-occurrences.

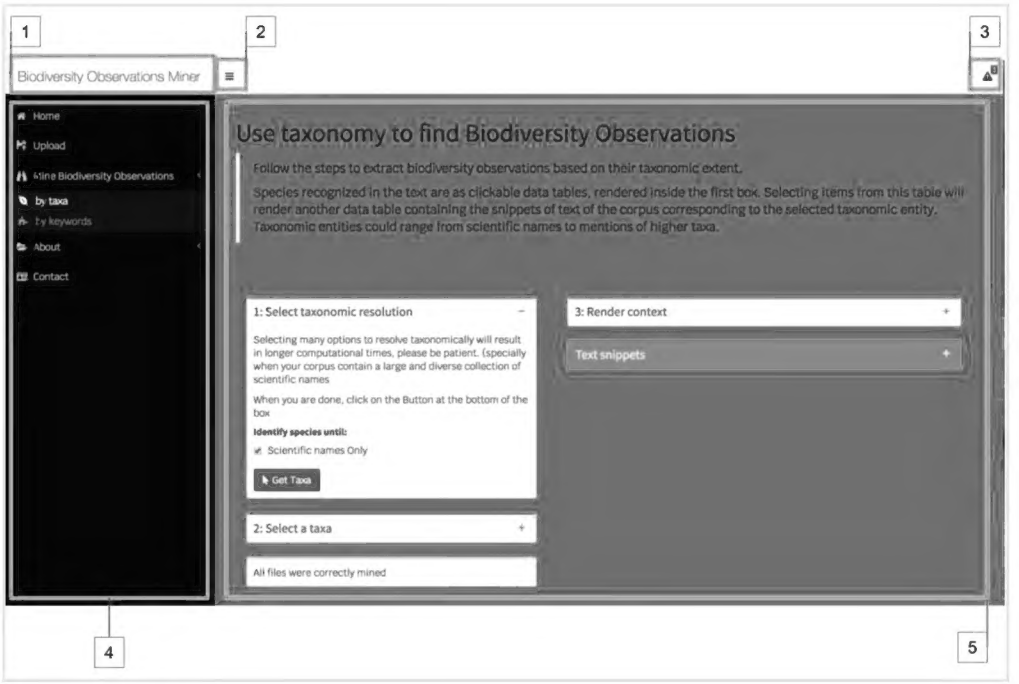

Figure 1. doi

Sections of Biodiversity Observations Miner (BOM) user interface: The figure illustrates the different parts that compose the user interface of BOM web application. The interface is composed of three main components, a header (white bar on top), a sidebar menu (dark blue at in the left side) and the main page (cyan in the centre). The header includes the application name (1), a button to collapse the sidebar menu (2) and a notification menu (3). The sidebar menu (4) contains the individual tabs to navigate across the functionalities of BOM. The main page (5) allows the setting of parameters and obtaining the results of the mining steps. In the main page, the header of setting type boxes are colour-coded yellow whereas the result boxes (i.e. Text snippets) are colour-coded with red headers.

Functional description:

OCR of PDF files

Before using Biodiversity Observations Miner, a user needs to create a corpus of relevant literature, stored as a collection of individual PDF files. This biodiversity literature corpus can be compiled by downloading PDFs of scientific articles from web databases such as Web of Science and Google Scholar. The collection of PDF files can be uploaded in batch to BOM. PDF versions from different publications can be very heterogeneous in nature. As such, plain text from PDF file(s) is recognised with the Google Tesseract tool for Optical Character Recognition (OCR) (Smith 2007). The Tesseract tool is a proven, well known, open-source OCR engine which can recognise many languages (Smith 2007). BOM performs the OCR of text with the Tesseract tool using the binding available in the scrapenames function from the taxize package (Chamberlain and Szöcs 2013). However, a portion of PDF files available in web databases does not come in machine-readable format. For example, digitised versions of old papers are usually stored as separate scanned images inside a single PDF file. Currently, BOM cannot handle this type of files and the user will be notified about the presence of such files in the literature corpus within the notification menu (*see* User's manual) (Suppl. material 1). For future updates of BOM, we will seek to include ways to automatically recognise and OCR all type of PDF files, including those with text stored as images.

Scientific name recognition

Biodiversity Observations Miner makes use of the Global Names Recognition and Discovery (GNRD) (Mozzherin et al. 2017) application programme interface (API) to recognise scientific names present of the OCR text. This API is part of the Global Names Architecture (GNA) (Pyle 2016), a name-based cyber-infrastructure which offers a set of open and free web services to find, index and organise biological scientific names (Mozzherin et al. 2017). It includes an algorithm (*biodiversity*) that parses scientific names from text with high accuracy (Mozzherin et al. 2017). Latin words, journal names or terms that resemble the Latin binomial structure of scientific names can cause confusions to the algorithm. However, errors in recognition are usually attributed to false positives rather than false negatives (Mozzherin et al. 2017). A current drawback of the *biodiversity* algorithm is that common names of species are not recognised in the corpus text. BOM includes a search option for taxonomic identification at higher taxonomic ranks (i.e. Family and Class) of the species names recognised in the text. This information is retrieved by querying the National Center of Biotechnology Information (NCBI) taxonomic database using the E-utilities RESTful API of NCBI. Functions to connect to both APIs are implemented in the R package taxize (Chamberlain and Szöcs 2013).

Calculating word co-occurrences

Individual sentences across the whole literature corpus are considered as text snippets that potentially contain one or more biodiversity observations of particular interest for a user of BOM. As such, word co-occurrence patterns can provide useful information to characterise the content of these text snippets. For example, the words "body" + "size" can be used to tag individual text snippets with information on allometric relations, functional trait relationships etc. In BOM, text strings from the literature corpus are split into sentences using a sentence tokeniser. Then, the individual elements (e.g. nouns, verbs, articles) of these sentences are annotated with a pre-trained, English based, natural language processing (NLP) model (Straka and Straková 2017). Finally, a skip-n-gram model is applied to the pool of tokenised sentences.

The skip-n-gram model is a practical, powerful model to infer context from text and is usually applied in processes such as speech recognition (Silge and Robinson 2016, Thessen et al. 2012). The value of "n" in the model defines the size (i.e. number of words) of the moving window applied to find word vectors in continuous text. These word vectors are constructed by selecting word pairs composed of a fixed word and all other possible combinations of words inside the moving window (Fig. 2). Word pairs are pooled together disregarding the individual distances between the fixed word and the other words inside the moving window. In BOM, a n = 6 was considered to construct the skip-n-gram model and we only included nouns, verbs, and adjectives into the moving window. This was done to prune common stop words (e.g. "the", "all", "and", "however") for co-occurrence calculations. The udpipe (Straka and Straková 2017) package for R was used for sentence tokenization, annotation and to apply the skip-n-gram model. Word co-occurrences are sorted by frequency counts before being presented to BOM users.

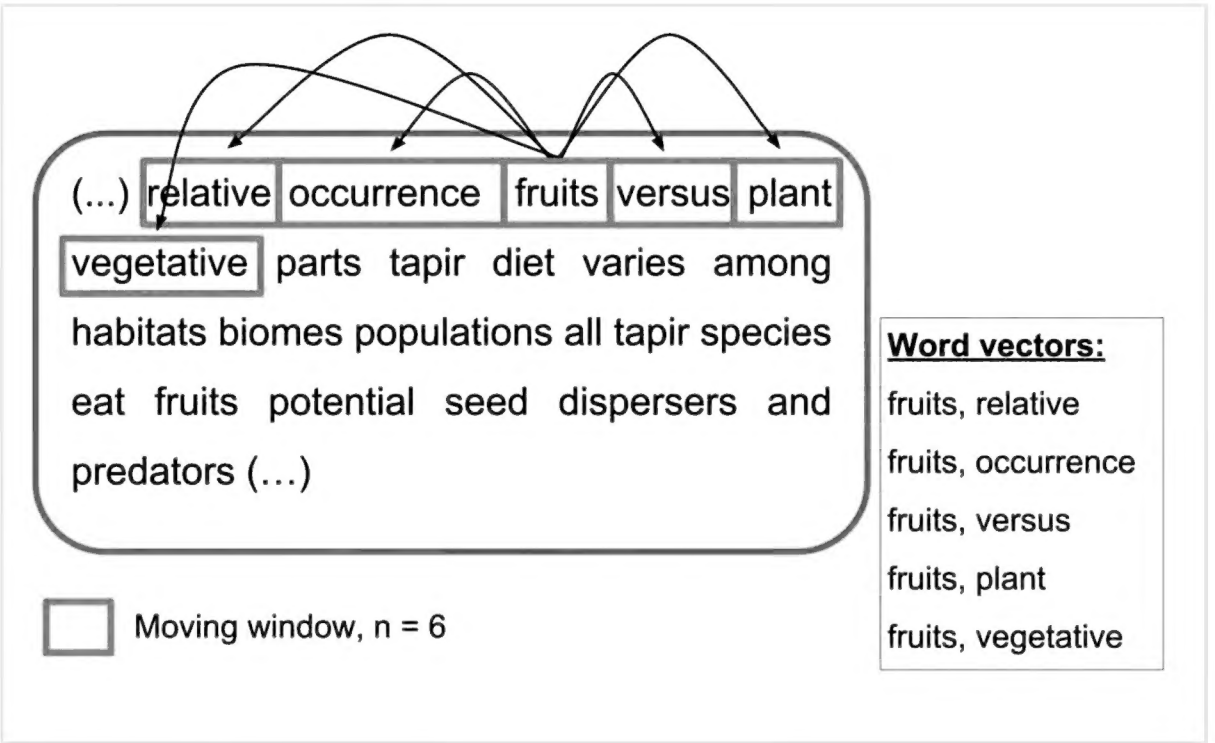

Figure 2. doi

Example of a moving window of n = 6 of a skip-n-gram model over a piece of text from O'Farrill et al. (2013). The text has been cleaned of common stop words (e.g. "the", "all", "however"). Inside the moving window, a central word is fixated (randomly) and all possible word pairs are considered as word vectors. After this step is completed, the moving window advances one word and repeats the process again. Frequencies of co-occurrences within the pool of word vectors are further used to rank word pairs.

Retrieving text snippets

BOM uses indexed scientific names and word co-occurrences to retrieve text snippets across all the uploaded literature corpus. This allows rapid discovery of targeted biodiversity observations inside the corpus text. First, with the **byTaxa** tab, the use of scientific names to retrieve text snippets and word co-occurences to characterise its content allows for rapid screening of literature based on the particular taxonomic interest of an individual user. Second, with the **byKeywords** tab, BOM also allows the retrieval of text snippets based on individual word co-occurrences only. These word co-occurrences can be further filtered using custom biodiversity dictionaries.

Biodiversity dictionaries

A biodiversity dictionary is a list of common terms used to describe a particular biodiversity observation. Currently, BOM lists biodiversity dictionaries matching text observations of frugivory and pollination, i.e. specific biotic interaction types. For example, the written description of a plant-animal interaction of frugivory might include terms such as *fruit*, *eat*, *disperse*, *swallow*, etc. (Fig. 3). Terms included in those biodiversity dictionaries were manually selected from a unigram term-frequency matrix created from sample articles known to contain biodiversity observations on frugivory or pollination. In creating these dictionaries, we limited the length of terms composing the biodiversity dictionary by discussing the rationale behind each term and eliminating ambiguous terms that might

match a large number of false positive snippets (i.e. snippets containing non-relevant information). However, because of the intrinsic heterogeneity of natural language to store biodiversity information, certain terms might match other type of observations. However, in BOM, terms in the biodiversity dictionaries are used to optionally filter the list word co-occurrences and not to index text snippets *per se*. This allows the user to finally determine if a particular combination of co-occurring words might lead to snippets containing information of interest (e.g. "eat" + "fruit" = *frugivory* whereas "eat" + "prey" = *predation*). In future updates, we aim to include more biodiversity dictionaries in the web version of BOM. Nevertheless, users running the application locally can also easily integrate custom biodiversity dictionaries of their own (see User's manual: Suppl. material 1).

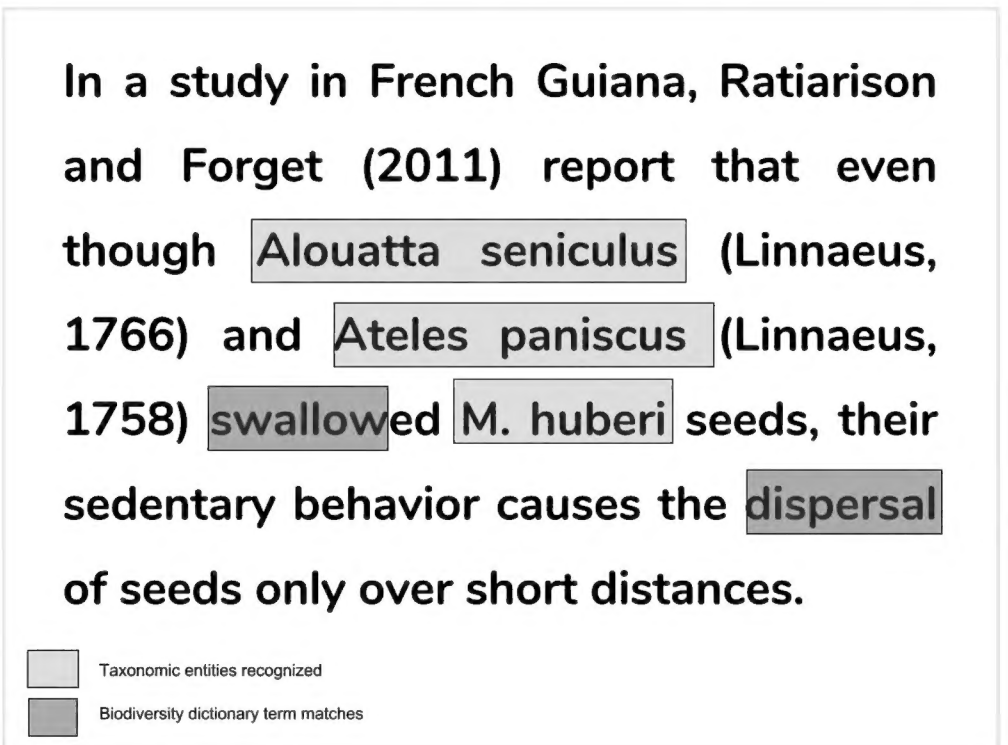

Figure 3. doi

Example of one text snippet resulting from running Biodiversity Observations Miner with O'Farrill et al. (2013) as input. This text snippet (*i.e. biodiversity observation*) contains data about a **frugivory** interaction between plants and animals. Here, biodiversity data comes from the description of the monkeys *Alouatta seniculus* and *Ateles paniscus* being frugivores of *M. huberi* fruits. The terms "swallow" and "dispersal" were part of the frugivory biodiversity dictionary included in BOM. Red boxes highlight the taxonomical entities recognised using the Global Names Architecture API implemented with the taxize (Chamberlain and Szöcs 2013) R package. The green boxes show the matches of frugivory dictionary terms within the text snippet.

Web location (URIs)

Homepage: https://fgabriel1891.github.io/BiodiversityObservationsMiner/

Download page: https://fgabriel1891.github.io/BiodiversityObservationsMiner/

Bug database: https://github.com/fgabriel1891/BiodiversityObservationsMiner/issues/

Technical specification

Platform: shiny, R.

Programming language: R

Operational system: Windows, OSx, Linux

Interface language: shiny-dashboard, shiny

Repository

Type: Git

Browse URI: BiodiversityObservationsMiner

Usage rights

Use license: Other

Implementation

Implements specification

Published literature in ecology holds a vast amount of information from centuries of research (Miller et al. 2012, Nunez-Mir et al. 2016). However, digitally storing this knowledge as text, in PDF files, limits its openness and accessibility. Thus, as Ecology moves towards a data-driven science (Michener et al. 2011, Petrič and Cestnik 2014, Senderov et al. 2017), the need for easy and standard access to biodiversity data increases (Edwards 2000, Michener and Jones 2012, Kissling et al. 2014, Kissling et al. 2018). Although recent publication practices are increasing the mobility and discoverability of biodiversity data (e.g. Wilkinson et al. 2016), finding information from literature can become challenging and time-consuming. In this sense, Biodiversity Observations Miner is a piece of software which contributes to the discovery, mobilisation and reuse of ecological data stored in scientific literature. BOM can be implemented inside biodiversity research workflows to filter candidate studies in meta-analysis, to discover biodiversity observations for testing hypothesis and to populate global-scale standard biodiversity databases like the Plant Trait Database (TRY: www.try-db.org) (Kattge et al. 2011), the Data Observation Network for Earth (DataOne: www.dataone.org) (Michener et al. 2011) or Global Biotic Interactions (GloBi: www.globalbioticinteractions.org).

In ecology and biodiversity science, computational methods such as machine learning algorithms have slowly integrated into research frameworks when compared with other

disciplines (Thessen 2016). Within the field of biodiversity data discovery, recent developments are making substantial progress to bridge this computational gap in ecology (Edwards 2000, Pyle 2016, Garnier et al. 2017, Mozzherin et al. 2017, Senderov et al. 2017, Senderov et al. 2018). As such, the use of proven algorithms through APIs and the open access of digital infrastructures such as the GNRD (Pyle 2016) will certainly foster future open software developments and digital workflows directed towards all research stages in ecology and biodiversity science. Text mining biodiversity observations of species functional traits and biotic interactions is particularly promising and can serve as a starting point to fill knowledge gaps that limit the advancement of ecology and biogeography as a science (Hortal et al. 2015).

The heterogeneity on terminologies describing particular biodiversity observations creates a challenge to automatically characterise text-based observations into standardised biodiversity data. Currently, there is a lack of standard terminologies to describe particular biodiversity observations. For instance, the term "eat" might match the textual description of many forms of biotic interactions (e.g. predation, frugivory, commensalism). We believe that initiatives, such as BOM, can benefit from future work that promotes the standardisation of terms via ontologies and controlled vocabularies. Furthermore, this could be further expanded to increase biodiversity dictionaries to match observations of natural history (e.g. dispersal distances, habitat preferences), biotic interactions (e.g. parasitism) or species functional traits (e.g. leaf area, flower phenology, body mass, wing length, mandible type, lifetime reproductive output) (Cornelissen et al. 2003, Moretti and Legg 2009, Kissling et al. 2018).

Audience

The target audience for this web application includes ecologists and biodiversity scientists at all career stages. Additionally, this application invites developers (ecologists or not) to suggest ideas for improvement. We are open to discussing additional ideas or new tools to expand the current functionalities of this web application.

Additional information

Dependencies

Biodiversity Observations Miner was written in R (R Development Core Team 2015) using the shiny (Chang et al. 2017) R package. Application user interface (UI) was built using the shiny-dashboard R package (Chang and Borges Ribeiro 2018).

Biodiversity Observations Miner makes use of R packages designed for text mining and base R functions. The taxize package is used to establish the API connection to the Global Names Recognition and Discovery (GNRF) tool. Taxize is also used for Optical Character Recognition (OCR) of the text in the PDFs and is done by GNA using the Google Tesseract Tool. The stringr is used for string manipulation. Details on the code and custom functions written for this application can be found in the GitHub Repository of this application. In

addition, BOM requires the following R packages to run locally: *shiny*, *shinydashboard*, *stringi*, *stringr, taxize, reshape, udpipe, tibble, DT*.

Acknowledgements

Biodiversity Observations Miner uses tools from GNA (GlobalNamesArchitecture) implemented in the taxize package. Thanks to Scott Chamberlain for modifications to the scrapenames function in taxize so it returns the OCR content of PDF files (https://github.com/ropensci/taxize/issues/614). Credits to the developers of the individual packages which is Biodiversity Observations Miner-dependent. Terms composing the pollination biodiversity dictionary were selected in collaboration with Joan Casanelles. Tomas Medina provided grammatical corrections and feedback for the first draft text.

Author contributions

GM developed Biodiversity Observations Miner with guidance, comments and input from WDK and EvL. GM wrote the first draft of the manuscript and WDK and EvL provided input. Terms composing the frugivory interactions dictionary were discussed between GM and WDK.

References

- Chamberlain SA, Szöcs E (2013) taxize: taxonomic search and retrieval in R. F1000Research 2: 191. https://doi.org/10.12688/f1000research.2-191.v2
- Chang W, Cheng J, Allaire J, Xie Y, McPherson J (2017) shiny:Web Application Framework for R. 1.0.5. CRAN. URL: https://CRAN.R-project.org/package=shiny
- Chang W, Borges Ribeiro B (2018) shiny dashboard: Create Dashboards with Shiny. 0.7.0. CRAN. URL: https://CRAN.R-project.org/package=shinydashboard
- Cornelissen J, Lavorel S, Garnier E, Diaz S, Buchmann N, Gurvich D, Reich P, Ter Steege H, Morgan H, Van Der Heijden M (2003) A handbook of protocols for standardised and easy measurement of plant functional traits worldwide. Australian Journal of Botany 51: 335-380. https://doi.org/10.1071/BT02124
- Edwards J (2001) The Global Biodiversity Information Facility: An international network of interoperabel biodiversity databases. Joho Chishiki Gakkaishi 10 (4): 58-61. https://doi.org/10.2964/jsik_kj00001039357
- Edwards JL (2000) Interoperability of biodiversity databases: Biodiversity information on every desktop. Science 289 (5488): 2312-2314. https://doi.org/10.1126/science.289.5488.2312
- Garnier E, Stahl U, Laporte M, Kattge J, Mougenot I, Kuehn I, Laporte B, Amiaud B, Ahrestani F, Boenisch G, Bunker D, Cornelissen J, Diaz S, Enquist B, Gachet S, Jaureguiberry P, Kleyer M, Lavorel S, Maicher L, Perez-Harguindeguy N, Poorter H, Schildhauer M, Shipley B, Violle C, Weiher E, Wirth C, Wright I, Klotz S (2017) Towards

a thesaurus of plant characteristics: an ecological contribution. Journal of Ecology 105: 298-309. https://doi.org/10.1111/1365-2745.12698

- Hearst M (2012) Text Data Mining. Oxford Handbooks Online https://doi.org/10.1093/oxfordhb/9780199276349.013.0034
- Hortal J, Bello Fd, F. Diniz-Filho JA, Lewinsohn T, Lobo J, Ladle R (2015) Seven Shortfalls that Beset Large-Scale Knowledge of Biodiversity. Annual Review of Ecology, Evolution, and Systematics 46 (1): 523-549. https://doi.org/10.1146/annurev-ecolsys-112414-054400
- Kattge J, Diaz S, Lavorel S, Prentice IC, Leadley P, Bönisch G, Garnier E, Westoby M, Reich P, Wright I (2011) TRY-a global database of plant traits. Global change biology 17 (9): 2905-2935. https://doi.org/10.1111/j.1365-2486.2011.02451.x
- Kissling WD, Dalby L, Fløjgaard C, Lenoir J, Sandel B, Sandom C, Trøjelsgaard K, Svenning J (2014) Establishing macroecological trait datasets: digitalization, extrapolation, and validation of diet preferences in terrestrial mammals worldwide. Ecology and Evolution 4 (14): 2913-2930. https://doi.org/10.1002/ece3.1136
- Kissling WD, Hardisty A, García EA, Santamaria M, Leo FD, Pesole G, Freyhof J, Manset D, Wissel S, Konijn J, Los W (2015) Towards global interoperability for supporting biodiversity research on essential biodiversity variables (EBVs). Biodiversity 16: 99-107. https://doi.org/10.1080/14888386.2015.1068709
- Kissling WD, Walls R, Bowser A, Jones MO, Kattge J, Agosti D, Amengual J, Basset A, van Bodegom PM, Cornelissen JHC, Denny EG, Deudero S, Egloff W, Elmendorf SC, Alonso García E, Jones KD, Jones OR, Lavorel S, Lear D, Navarro LM, Pawar S, Pirzl R, Rüger N, Sal S, Salguero-Gómez R, Schigel D, Schulz K, Skidmore A, Guralnick RP (2018) Towards global data products of Essential Biodiversity Variables on species traits. Nature ecology & evolution 2 (10): 1531-1540. https://doi.org/10.1038/s41559-018-0667-3
- Michener W, Vieglais D, Vision T, Kunze J, Cruse P, Janée G (2011) DataONE: Data Observation Network for Earth - preserving data and enabling innovation in the biological and environmental sciences. D-Lib Magazine 17 https://doi.org/10.1045/january2011-michener
- Michener W, Jones M (2012) Ecoinformatics: supporting ecology as a data-intensive science. Trends in Ecology & Evolution 27 (2): 85-93. https://doi.org/10.1016/j.tree.2011.11.016
- Miller J, Dikow T, Agosti D, Sautter G, Catapano T, Penev L, Zhang Z, Pentcheff D, Pyle R, Blum S, Parr C, Freeland C, Garnett T, Ford LS, Muller B, Smith L, Strader G, Georgiev T, Bénichou L (2012) From taxonomic literature to cybertaxonomic content. BMC Biology 10 (1): 87. https://doi.org/10.1186/1741-7007-10-87
- Moretti M, Legg C (2009) Combining plant and animal traits to assess community functional responses to disturbance. Ecography 32 (2): 299-309. https://doi.org/10.1111/j.1600-0587.2008.05524.x
- Mozzherin D, Myltsev A, Patterson D (2017) “gnparser”: a powerful parser for scientific names based on Parsing Expression Grammar. BMC Bioinformatics 18: 279. https://doi.org/10.1186/s12859-017-1663-3
- Nguyen NH, Soto A, Kontonatsios G, Batista-Navarro R, Ananiadou S (2017) Constructing a biodiversity terminological inventory. PLOS ONE 12 (4): e0175277. https://doi.org/10.1371/journal.pone.0175277

- Nunez-Mir G, Iannone B, Pijanowski B, Kong N, Fei S (2016) Automated content analysis: addressing the big literature challenge in ecology and evolution. Methods in Ecology and Evolution 7 (11): 1262-1272. https://doi.org/10.1111/2041-210x.12602
- O'Farrill G, Galleti M, Campos-Arceiz A (2013) Frugivory and seed dispersal by tapirs: an insight on their ecological role. Integrative Zoology 8 (1): 4-17. https://doi.org/10.1111/j.1749-4877.2012.00316.x
- Petrič I, Cestnik B (2014) Predicting future discoveries from current scientific literature. Methods in Molecular Biology. https://doi.org/10.1007/978-1-4939-0709-0_10
- Poelen J, Simons J, Mungall C (2014) Global biotic interactions: An open infrastructure to share and analyze species-interaction datasets. Ecological Informatics 24: 148-159. https://doi.org/10.1016/j.ecoinf.2014.08.005
- Pyle RL (2016) Towards a Global Names Architecture: The future of indexing scientific names. ZooKeys 550: 261-281. https://doi.org/10.3897/zookeys.550.10009
- R Development Core Team (2015) R Software for statistical computing. 3.4.3 (2017-11-30) - "Kite-Eating Tree".
- Saffer J, Burnett V (2014) Introduction to biomedical literature text mining: context and objectives. Methods in Molecular Biology. https://doi.org/10.1007/978-1-4939-0709-0_1
- Sautter G, Böhm K, Agosti D (2006) A combining approach to find all taxon names (FAT). Biodiversity Informatics 3: 46-58. https://doi.org/10.17161/bi.v3i0.34
- Senderov V, Georgiev T, Agosti D, Catapano T, Sautter G, Tuama ÉÓ, Franz N, Simov K, Stoev P, Penev L (2017) OpenBiodiv: an implementaion of a semantic system running on top of the biodiversity knowledge graph. Proceedings of TDWG 1: e20084. https://doi.org/10.3897/tdwgproceedings.1.20084
- Senderov V, Simov K, Franz N, Stoev P, Catapano T, Agosti D, Sautter G, Morris RA, Penev L (2018) OpenBiodiv-O: ontology of the OpenBiodiv knowledge management system. Journal of Biomedical Semantics 9: 5. https://doi.org/10.1186/s13326-017-0174-5
- Silge J, Robinson D (2016) tidytext: Text Mining and Analysis Using Tidy Data Principles in R. The Journal of Open Source Software 1 (3): 37. https://doi.org/10.21105/joss.00037
- Smith R (2007) An Overview of the Tesseract OCR Engine . *ICDAR 2007. Ninth International Conference*., Vol. 2, pp. 629-633).. *Document Analysis and Recognition*.
- Straka M, Straková J (2017) Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies. https://doi.org/10.18653/v1/k17-3009
- Tari L, Patel J (2014) Systematic drug repurposing through text mining. Methods in Molecular Biology. https://doi.org/10.1007/978-1-4939-0709-0_14
- Thessen A, Parr CS (2014) Knowledge extraction and semantic annotation of text from the Encyclopedia of Life. PLoS ONE 9 (3): e89550. https://doi.org/10.1371/journal.pone.0089550
- Thessen AE, Cui H, Mozzherin D (2012) Applications of natural language processing in biodiversity science. Advances in Bioinformatics 2012: 391574. https://doi.org/10.1155/2012/391574
- Thessen AE (2016) Adoption of machine learning techniques in Ecology and Earth Science. One Ecosyste https://doi.org/10.7287/peerj.preprints.1720
- Wilkinson MD, Dumontier M, Aalbersberg IJJ, Appleton G, Axton M, Baak A, Blomberg N, Boiten J-W, da Silva Santos LB, Bourne PE, Bouwman J, Brookes AJ, Clark T,

Crosas M, Dillo I, Dumon O, Edmunds S, Evelo CT, Finkers R, Gonzalez-Beltran A, Gray AJG, Groth P, Goble C, Grethe JS, Heringa J, 't Hoen PAC, Hooft R, Kuhn T, Kok R, Kok J, Lusher SJ, Martone ME, Mons A, Packer AL, Persson B, Rocca-Serra P, Roos M, van Schaik R, Sansone S-A, Schultes E, Sengstag T, Slater T, Strawn G, Swertz MA, Thompson M, van der Lei J, van Mulligen E, Velterop J, Waagmeester A, Wittenburg P, Wolstencroft K, Zhao J, Mons B (2016) The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data 3: 160018. https://doi.org/10.1038/sdata.2016.18

Supplementary material

Suppl. material 1: BOM_USER_MANUAL doi

Authors: Gabriel Muñoz
Data type: user's manual
Brief description: Biodiversity Observation User's manual. Follow this guide to upload literature and mine biodiversity observations using BOM.
Filename: BOM_MANUAL.pdf - Download file (938.02 kb)